

## Automatische Inhaltsererschließung in der Fachinformation: eine Evaluation zur maschinellen Indexierung sozialwissenschaftlicher Forschungsliteratur

Kempf, Andreas Oskar

Veröffentlichungsversion / Published Version

Zeitschriftenartikel / journal article

*Dieser Beitrag ist mit Zustimmung des Rechteinhabers aufgrund einer (DFG geförderten) Allianz- bzw. Nationallizenz frei zugänglich. / This publication is with permission of the rights owner freely accessible due to an Alliance licence and a national licence (funded by the DFG, German Research Foundation) respectively.*

### Empfohlene Zitierung / Suggested Citation:

Kempf, A. O. (2013). Automatische Inhaltsererschließung in der Fachinformation: eine Evaluation zur maschinellen Indexierung sozialwissenschaftlicher Forschungsliteratur. *Information - Wissenschaft und Praxis*, 64(2-3), 96-106.  
<https://doi.org/10.1515/iwp-2013-0011>

### Nutzungsbedingungen:

Dieser Text wird unter einer Deposit-Lizenz (Keine Weiterverbreitung - keine Bearbeitung) zur Verfügung gestellt. Gewährt wird ein nicht exklusives, nicht übertragbares, persönliches und beschränktes Recht auf Nutzung dieses Dokuments. Dieses Dokument ist ausschließlich für den persönlichen, nicht-kommerziellen Gebrauch bestimmt. Auf sämtlichen Kopien dieses Dokuments müssen alle Urheberrechtshinweise und sonstigen Hinweise auf gesetzlichen Schutz beibehalten werden. Sie dürfen dieses Dokument nicht in irgendeiner Weise abändern, noch dürfen Sie dieses Dokument für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, aufführen, vertreiben oder anderweitig nutzen.

Mit der Verwendung dieses Dokuments erkennen Sie die Nutzungsbedingungen an.

**gesis**  
Leibniz-Institut  
für Sozialwissenschaften

### Terms of use:

This document is made available under Deposit Licence (No Redistribution - no modifications). We grant a non-exclusive, non-transferable, individual and limited right to using this document. This document is solely intended for your personal, non-commercial use. All of the copies of this documents must retain all copyright information and other information regarding legal protection. You are not allowed to alter this document in any way, to copy it for public or commercial purposes, to exhibit the document in public, to perform, distribute or otherwise use the document in public.

By using this particular document, you accept the above-stated conditions of use.

Mitglied der  
**Leibniz**  
Leibniz-Gemeinschaft

Andreas Oskar Kempf, Köln

# Automatische Inhaltserschließung in der Fachinformation

## Eine Evaluation zur maschinellen Indexierung sozialwissenschaftlicher Forschungsliteratur

Der Artikel basiert auf einer Masterarbeit mit dem Titel „Automatische Indexierung in der sozialwissenschaftlichen Fachinformation. Eine Evaluationsstudie zur maschinellen Erschließung für die Datenbank SOLIS“ (Kempf 2012), die im Rahmen des Aufbaustudiengangs Bibliotheks- und Informationswissenschaft an der Humboldt-Universität zu Berlin am Lehrstuhl Information Retrieval verfasst wurde. Auf der Grundlage des Schalenmodells zur Inhaltserschließung in der Fachinformation (vgl. Krause 1996, 2006) stellt der Artikel Evaluationsergebnisse eines automatischen Erschließungsverfahrens für den Einsatz in der sozialwissenschaftlichen Fachinformation vor. Ausgehend von dem von Krause beschriebenen Anwendungsszenario, wonach SOLIS-Datenbestände (Sozialwissenschaftliches Literaturinformationssystem) von geringerer Relevanz automatisch erschlossen werden sollten, wurden auf dieser Dokumentgrundlage zwei Testreihen mit der Indexierungssoftware MindServer der Firma Recommend<sup>1</sup> durchgeführt. Neben den Auswirkungen allgemeiner Systemeinstellungen in der ersten Testreihe wurde in der zweiten Testreihe die Indexierungsleistung der Software für die Rand- und die Kernbereiche der Literaturdatenbank miteinander verglichen. Für letztere Testreihe wurden für beide Bereiche der Datenbank spezifische Versionen der Indexierungssoftware aufgebaut, die anhand von Dokumentkorpora aus den entsprechenden Bereichen trainiert wurden. Die Ergebnisse der Evaluation, die auf der Grundlage intellektuell generierter Vergleichsdaten erfolgt, weisen auf Unterschiede in der Indexierungsleistung zwischen Rand- und Kernbereichen hin, die einerseits gegen den Einsatz automatischer Indexierungsverfahren in den Randbereichen sprechen. Andererseits deutet sich an, dass sich die Indexierungsergebnisse durch den Aufbau fachteilgebietspezifischer Trainingsmengen verbessern lassen.

**Deskriptoren:** Automatische Indexierung, Fachinformation, Sozialwissenschaften, Thesaurus, Bewertung

### **Automatic indexing of domain-specific information. An evaluation of automated content cataloguing of social science research literature**

This article is based on a Master thesis with the title “Automatische Indexierung in der sozialwissenschaftlichen Fachinformation. Eine Evaluationsstudie zur maschinellen Erschließung für die Datenbank SOLIS” (Kempf 2012) written within the framework of the postgraduate study program Library and Information Science at Humboldt-Universität zu Berlin at the chair of Information Retrieval. On the basis of the so-called ‘Shell Model’ (Krause 1996, 2006) for domain-specific content cataloguing it presents evaluation results of an automatic indexing tool for cataloguing of social science research literature. Taking the concrete application scenario formulated by Krause, which suggests that SOLIS-data (Social Science Literature Information System) of less relevance should be indexed automatically, the software MindServer by Recommend was tested in two test series on exactly this data. While in the first test series the system’s general settings were tested in the second test series the indexing performance for key and for border areas of the database were compared. For this purpose, sub-discipline-specific versions of the software were built up, which were trained on the basis of corresponding data corpora. The results, evaluated on the basis of intellectually generated comparative data, indicate differences in the quality of indexing for key and for border areas of the database which on the one hand speak against the use of automatic indexing for this area of the database. On the other hand the tests suggest that by building up sub-discipline-specific corpora of training the indexing results could be improved.

**Keywords:** automatic indexing, domain-specific information, ‘Shell Model’, thesaurus, evaluation

<sup>1</sup> [www.recommend.com](http://www.recommend.com)



## L'indexation automatique dans l'information spécialisée. Une évaluation du catalogage automatisé de la littérature de recherche en sciences sociales

Cet article est basé sur un mémoire de Master II intitulé « Automatische Indexierung in der sozialwissenschaftlichen Fachinformation. Eine Evaluationsstudie zur maschinellen Erschließung für die Datenbank SOLIS » (Kempf 2012) rédigé dans le cadre du cursus post-gradué Bibliothéconomie et Sciences de l'Information de Humboldt-Universität zu Berlin à la chaire Information Retrieval. Se basant sur le modèle dit « modèle des strates » (Krause 1996, 2006) pour le catalogage d'un contenu spécifique à un domaine, il présente des résultats d'une étude qui porte sur les outils d'indexation automatique dans la littérature de recherche en sciences sociales. En partant du scénario concret formulé par Krause qui stipule que les données SOLIS (Système d'Information en Littérature dans les Sciences Sociales) d'importance moindre devraient être indexées de manière automatisée, le logiciel MindServer fait par Recommind a été testé dans deux séries de tests portant exactement sur ces données. Tandis que dans la première série les paramètres généraux ont été testés, la deuxième série portait sur la performance dans le domaine de l'indexation de données centrales et périphériques. A cet effet, on a établi des versions sous-spécifiques du logiciel qui étaient entraînées sur des corpus de données correspondant aux sous-disciplines. Les résultats, évalués sur les bases de données comparatives générées intellectuellement, indiquent des différences dans la qualité d'indexation pour les données centrales et périphériques de la banque de données qui mettent en garde contre l'usage de l'indexation automatique dans cette partie de la banque de données. De l'autre côté, les tests révèlent qu'en établissant des corpus sous-spécifiques d'entraînement les résultats d'indexation peuvent être améliorés.

**Mots-clés:** indexation automatique, information spécialisée, modèle des strates, thésaurus, évaluation

## 1 Einleitung

Mit Zunahme der digitalen Verfügbarkeit von Metadaten und Volltexten werden in jüngerer Zeit automatische Indexierungsverfahren als eine mögliche Antwort auf den enormen Ressourcenaufwand bei der Inhaltserschließung diskutiert. Dabei reicht der Einsatz automatischer Erschließungsverfahren für einen klassischen Anwen-

dungsbereich der dokumentarischen Profession, die Medien- bzw. Pressedokumentation, sehr viel weiter zurück. Auf der Grundlage schmalere Thesauri und Klassifikationen lieferten die angewendeten Verfahren schnelle Erfolge und eine deutliche Kosten- und Zeitersparnis.<sup>2</sup>

Mit Umstieg auf Online-Kataloge in den 1990er Jahren setzte auch in den Bibliotheken eine Diskussion um die bisherige Form der Inhaltserschließung ein. Wichtige Initiativen zur Entwicklung und Anwendung computerunterstützter Inhaltserschließung, die sowohl auf linguistischen als auch auf statistischen Verfahren beruhen, bilden in der deutschsprachigen Bibliothekswelt die Projekte MILOS I und II, das Nachfolgeprojekt KASKADE sowie die Projekt OSIRIS (vgl. Siegmüller 2007) und PETRUS (vgl. Schöning-Walter 2011 sowie DNB 2010).

Seit etwa zehn Jahren sammeln auch eine Reihe von Fachinformationszentren Erfahrungen auf dem Gebiet der automatischen Erschließung. Mit dem Auftrag, die jeweilige Fachgemeinschaft mit wissenschaftlicher Information zu versorgen, leisten sie in besonderer Weise eine zeitnahe Strukturierung und tiefgehende Aufbereitung von Informationen. Ein Beispiel für eine bereits implementierte prozessunterstützte Erschließung auf der Grundlage einer vor allem computerlinguistisch operierenden Indexierungssoftware bildet die Dokumentation psychologischer Literatur und Medien in der Datenbank PSYINDEX am Zentrum für Psychologische Information und Dokumentation (ZPID) (vgl. Gerards et al. 2006 sowie Gerards 2011, zur Unterscheidung der verschiedenen Verfahren siehe weiter unten). In den Dokumentationsablauf integriert, liefert die automatische Indexierungssoftware auf der Basis von Dokumenttitel, Abstract und von Autoren angegebenen Stichwörtern Deskriptorvorschläge zur Unterstützung der intellektuellen Inhaltserschließung.<sup>3</sup>

Mit dem Ziel, durch ein semiautomatisches Indexierungsverfahren Ressourcen freizusetzen, beschäftigt sich in jüngerer Zeit auch GESIS – Leibniz-Institut für Sozialwissenschaften mit dem Thema der softwareunterstützten Inhaltserschließung. Ausschlag gab die Anschaffung der Software MindServer zur Entwicklung eines *Search Term*

<sup>2</sup> Jüngere Beispiele mit mittlerweile deutlich umfangreicheren Kategorienschemata bilden etwa die Nachrichtenagentur Reuters sowie der Verlag Gruner + Jahr und das ZDF (vgl. Bertram 2005 sowie exemplarisch Lingelbach-Hupfauer/Laute 2009, Lingelbach-Hupfauer 2011).

<sup>3</sup> Daneben befassen sich sowohl das Deutsche Institut für Internationale Pädagogische Forschung (DIPF) (vgl. Wissel 2011) als auch seit Längerem die Zentralbibliothek für Wirtschaftswissenschaften (ZBW) mit automatischen Indexierungsverfahren (vgl. Groß 2010, Groß/Faden 2010).



*Recommenders* für das Fachportal sowieport im Rahmen eines DFG-geförderten Projektes zu Mehrwertdiensten für das Information Retrieval (IRM) (vgl. Mayr et al. 2009).

Die Absicht, teilweise automatische Erschließungsverfahren für die sozialwissenschaftliche Literaturdokumentation zu nutzen, geht allerdings bereits auf das sog. Schalenmodell des früheren IZ-Präsidenten Jürgen Krause (1996, 2006) zurück, wonach die inhaltliche Erschließung der Literaturdatenbank SOLIS nach unterschiedlichen Niveaustufen denkbar sei, die gleichsam verschiedene Schalen um einen tief und qualitativ hochwertig erschlossenen Kernbereich bildeten. Die äußerste Schale schlägt Krause für eine automatische Indexierung vor.

Diese frühen Überlegungen bilden die Ausgangsbasis, um den Einsatz automatischer Indexierungsverfahren für die Literaturerschließung in SOLIS in einer ersten Testreihe allgemein und in einer zweiten Testreihe im Vergleich zwischen Kern- und Randbereichen der Datenbank anhand intellektuell erhobener Vergleichsdaten zu evaluieren. Nach einer Kurzeinführung in die Inhaltsererschließung, einen Überblick über die verschiedenen Verfahrensansätze und die Vorstellung der Evaluationskriterien wird im folgenden Beitrag in den konkreten Anwendungskontext eingeführt bevor abschließend die konkreten Ergebnisse der beiden Testreihen präsentiert werden.

## 2 Grundlagen intellektueller und maschineller Inhaltsererschließung

Intellektuelle und maschinelle Verfahren der Inhaltsererschließung weisen deutliche Unterschiede auf. Der intellektuelle Indexierungsvorgang gründet allgemein auf einem zweistufigen Prozess aus Inhaltsanalyse und Inhaltsdarstellung (vgl. Nohr 2004<sup>5</sup>). Mit dem Ziel, ein inhaltsbeschreibendes Dokument-Surrogat zu erstellen, wird der Inhalt eines Dokumentes zunächst intellektuell auf der Grundlage entsprechenden Kontextwissens verarbeitet und anschließend in eine Indexierungssprache übersetzt. Die Beschreibungsmerkmale eines Dokuments werden somit nicht dem Dokument selbst, sondern unterschiedlichen Dokumentationssprachen entnommen, im Fall der sozialwissenschaftlichen Fachinformation dem Thesaurus<sup>4</sup> sowie der Klassifikation Sozialwissenschaften<sup>5</sup>.

<sup>4</sup> <http://www.gesis.org/unser-angebot/recherchieren/thesauri-und-klassifikationen/thesaurus-sozialwissenschaften/>

<sup>5</sup> [http://www.gesis.org/fileadmin/upload/dienstleistung/tools\\_standards/Klassifikation\\_Sozialwissenschaften.pdf](http://www.gesis.org/fileadmin/upload/dienstleistung/tools_standards/Klassifikation_Sozialwissenschaften.pdf)

Während die intellektuelle Erschließung für die konsistente Wiedergabe des Dokumentinhaltes somit zum einen bis auf die Bedeutungsebene eines Textes vordringt, bewegen sich maschinelle Verfahren im Gegensatz dazu alleine auf der sprachlichen Oberfläche der Dokumente. Zum anderen steht bei der intellektuellen Erschließung vor allem die korrekte Repräsentation des Dokumentinhaltes im Vordergrund. Automatische Verfahren zielen hingegen sehr viel stärker auf die Wiederauffindbarkeit des Dokumentes (vgl. Nohr 2005<sup>3</sup>). Sie zeichnen sich daher in der Regel durch die Einbindung in ein sog. Best-Match-Retrieval aus. Die Ergebnisanzeige bildet die Ähnlichkeit der Treffer zur Suchanfrage entsprechend eines sog. Relevance Ranking ab. Intellektuelle Indexierungsverfahren sind hingegen traditionell zumeist in sog. Exact-Match-Retrieval-Systeme eingebunden. Der Abgleich zwischen Sucheintrag und Ergebnismenge basiert auf einer binären Unterscheidung zwischen Treffern, die exakt der formulierten Suchanfrage entsprechen und Nicht-Treffern, die die gesamte restliche Dokumentenmenge bilden.

Die Verlagerung von einem Exact- hin zu einem Best-Match-Retrieval spiegelt sich auch in der Diskussion um die Entwicklung und Einführung prozessunterstützender Erschließungsverfahren wider. Ausgehend von der Vielfalt sprachlicher Ausdrucksweisen führen Kritiker an, dass es unmöglich sei, diese adäquat über maschinelle Verfahren abzubilden. Das zentrale Argument hinter dieser Position lautet, dass es für die Inhaltsererschließung unerlässlich sei, bis auf die Bedeutungsebene sprachlicher Zeichen vorzudringen. Erst unter Einbezug des Kontextes, der über ein maschinelles Verfahren nur bedingt zugänglich sei, könne die inhaltliche Wiedergabe möglich sein. Demgegenüber argumentieren Anhänger automatischer Verfahren, dass unterschiedliche Untersuchungen durchaus Belege für eine hohe Indexierungsqualität der Verfahren im Sinne der Wiederauffindbarkeit der Dokumente erbracht hätten (vgl. ebd.).

## 3 Verfahrensansätze der automatischen Erschließung

Verfahren der automatischen Indexierung lassen sich nach unterschiedlichen Ansätzen unterscheiden. In Bezug auf die Ermittlung inhaltskennzeichnender Indexterme lässt sich zwischen statistischen und computerlinguistischen Verfahren unterscheiden. Für die sich daran anschließende Zuordnung von Indextermen aus einem kontrollierten Vokabular, wie sie beim sog. Additions- im



Gegensatz zum Extraktionsverfahren vorgenommen wird, stehen begriffsorientierte Verfahrensansätze. Die im Folgenden näher ausgeführten Verfahrensansätze werden vielfach miteinander kombiniert.

### 3.1 Statistische Verfahren

Statistische Verfahren stellen die ersten und daher am weitesten entwickelten Verfahrensansätze dar. Die zentrale Annahme, die diesen Verfahren zugrunde liegt, lautet, dass die Frequenz, in der ein Term in einem Dokument vorkommt, etwas über die Bedeutung dieses Terms aussagt. Neben der Auffassung, dass aus den vorhandenen Termen eines Dokumentes eine bestimmte Selektion getroffen werden muss, da sich nicht alle als Indexierungsterme eignen, wird ferner davon ausgegangen, dass die in dieser Form ausgewählten Indexierungsterme in unterschiedlicher Weise zu der Bedeutung eines Textes beitragen. Als zentrale Voraussetzung für den Einsatz eines Best-Match Retrieval-Verfahrens (siehe oben) werden die Terme somit unterschiedlich gewichtet.

### 3.2 Linguistische Verfahren

Computer- bzw. informationslinguistische Verfahren ermitteln Indexterme hingegen auf der Basis sprachlicher Gesetzmäßigkeiten. Unter Einbezug von Morphologie und Syntax zielen sie darauf ab, die Vielfalt sprachlicher Phänomene zu reduzieren (vgl. Bertram 2005). Zu den Textanalyseschritten gehören etwa die Eliminierung sog. Stopp-Wörter, u. a. Artikel oder Präpositionen, und die Reduktion von Wortformen etwa durch die Rückführung grammatikalischer Flexionsformen auf ihre Grundform (Lemmatisierung), die Zerlegung unterschiedlicher morphologischer Varianten auf ihre Stammform (Stemming) und die Zerlegung von Komposita.

In diesem Zusammenhang lässt sich zwischen regelbasierten und wörterbuchgestützten Ansätzen unterscheiden, die beide zumeist miteinander kombiniert werden. Regelbasierte Verfahren gründen bei ihrer Analyse auf sprachspezifischen Regeln, die in Form von Algorithmen aufgestellt werden. Über diese Regeln, die den Pflegeaufwand relativ gering halten, lassen sich allgemein gehaltene Vorschriften formulieren, die gleichzeitig für eine Vielzahl von Anwendungsbeispielen gelten. Wörterbuchbasierte Verfahren zeichnen sich hingegen dadurch aus, dass für die linguistische Analyse ein Wörterbuch hinterlegt ist und somit keine den Einzelfall übergreifenden Regeln aufgestellt werden. Diese Form der linguisti-

schen Analyse eignet sich etwa für das Deutsche aufgrund der Vielzahl möglicher Kompositabildungen besser als für das Englische. Gleichwohl zeichnen sich diese Verfahren durch einen hohen Pflegeaufwand aus.

### 3.3 Begriffsorientierte Verfahren

Begriffsorientierte Verfahren versuchen, anders als computerlinguistische Verfahren, die sich für die Vergabe geeigneter Indexterme ausschließlich auf der Zeichenebene bewegen, bis auf die semantische Ebene der Wörter vorzudringen. Grundlage hierfür bleibt wie bei allen maschinellen Indexierungsverfahren allerdings die sprachliche Oberfläche des Volltextes bzw. Abstracts. Auf der Basis einer Textwortanalyse werden bedeutungstragende Wörter mit einem zugrundeliegenden kontrollierten Vokabular abgeglichen, aus dem die entsprechenden Indexterme abgeleitet werden. Für die Disambiguierung und Zusammenführung verschiedener Benennungen eines Begriffs wird zumeist erneut sowohl auf statistische als auch informationslinguistische Verfahren zurückgegriffen. Die Implementierung eines solchen Verfahrens ist zum einen allerdings mit einem hohen Pflegeaufwand verbunden, da das hinterlegte Vokabular an Deskriptoren stets aktualisiert werden muss. Zum anderen ist die Indexierung tendenziell lückenhaft, da sie nur zeitversetzt erfolgen kann (vgl. Nohr 2004<sup>5</sup> sowie Siegmüller 2007).

## 4 Evaluationsformen automatischer Erschließung

Die Bewertung automatischer Erschließungsverfahren erfolgt zumeist über Retrieval-Tests. Angelehnt an das binär strukturierte Boolesche Retrieval, basiert die Bewertung in erster Linie auf einer zweistufigen Bewertungsskala aus relevanten und nicht-relevanten Indexierungsergebnissen (vgl. Womser-Hacker 2004<sup>5</sup>). Auf dieser Grundlage werden die beiden Standardmaße zur Messung der Effektivität eines Retrieval-Systems, *Recall* und *Precision* gebildet.

Der *Recall* gibt die Vollständigkeit eines Retrieval-Ergebnisses an. Er steht für das Verhältnis zwischen selektierten relevanten Dokumenten und in der Dokumentensammlung insgesamt vorhandenen relevanten Dokumenten. Die *Precision* ermittelt indessen die Genauigkeit eines Retrieval-Ergebnisses. Sie stellt das Verhältnis zwischen selektierten relevanten Dokumenten und der Gesamtanzahl nachgewiesener Dokumente dar (vgl. Stock/Stock 2008).

Beide Maße, die in einem Spannungsverhältnis zueinander stehen, ergänzen sich in ihrer Aussage zur Effektivität eines Retrieval-Systems. So trifft der *Recall*-Wert einzig eine Aussage darüber, wie vollständig ein Retrieval-Ergebnis ausfällt, ohne die Ballastquote an nicht-relevanten Dokumenten, die ebenso in der Trefferanzeige enthalten sein können, einzubeziehen. Im Gegensatz dazu gibt der *Precision*-Wert allein die Effektivität des Retrieval-Systems an, nicht-relevante Dokumente aus der Trefferanzeige herauszufiltern.

Im Folgenden wird das Retrieval-Ergebnis nicht in Bezug auf den gesamten Bestand einer Dokumentensammlung, sondern einzig die Indexierungsleistung und somit die Vergabe der Deskriptoren sowie die Zuordnung der Klassifikationsnotationen auf der Ebene der einzelnen Dokumente untersucht. Erschließungsleistung auf der einen und das Antwortverhalten eines Retrieval-Systems auf der anderen Seite sind gleichwohl auf das Engste miteinander verknüpft. Die Repräsentation des Dokumentinhaltes in Form von Deskriptoren und Klassifikation entscheidet über die Auffindbarkeit eines Dokumentes und somit über die Effektivität eines Retrieval-Systems.

Bei der Auswertung der automatisch generierten Erschließungsergebnisse wurde die manuelle Indexierung als Relevanzmaßstab angelegt. Für die nachfolgend beschriebenen Testläufe stellt der *Recall* somit das Verhältnis zwischen (automatisch) selektierten relevanten, d. h. ebenso manuell vergebenen, Deskriptoren bzw. Klassifikationsnotationen und im intellektuell generierten Indexat insgesamt aufgeführten Deskriptoren und Notationen dar. Die *Precision* hingegen steht für den Quotienten aus der Überschneidungsmenge der sowohl intellektuell als auch maschinell vergebenen Deskriptoren bzw. Klassifikationen und der Gesamtanzahl maschinell generierter Deskriptoren bzw. Klassifikationsnotationen.<sup>6</sup>

Angesichts der großen Treffermengen in gängigen Suchmaschinen kommt der Präzision und damit der Fähigkeit eines Retrieval-Systems, Ballast herauszufiltern, eine größere Bedeutung zu als der Vollständigkeit eines Suchergebnisses. In der vorliegenden Untersuchung wurde daher bei manchen Testläufen die Präzision des Indexierungsergebnisses – gemäß der manuellen Erschließung – an bestimmten Punkten, den sog. *cut-off* Levels,

gemessen, hierzu ließ sich der von der Indexierungssoftware ermittelte Konfidenzwert nutzen.

Daneben wurde die Indexierungskonsistenz, die im vorliegenden Fall den Grad an Übereinstimmung zwischen intellektueller und maschineller Erschließung misst, als Untersuchungskriterium aufgenommen. Sie wird aus dem Quotienten aus der Anzahl an Überschneidung zwischen beiden Indexierungsverfahren und der Gesamtzahl sämtlicher intellektuell sowie maschinell erzeugter Erschließungsvorschläge berechnet (vgl. Stubbs et al. 1999, Rolling 1981 sowie Stock/Stock 2008).<sup>7</sup>

## 5 Datengrundlage und Erschließungsinstrumente

Die Datenbank SOLIS stellt mit mehr als 400.000 Datensätzen in Form von bibliographischen und inhaltlichen Angaben sowohl zu Monographien und Sammelwerken als auch zu Zeitschriftenaufsätzen, Sammelwerksbeiträgen und Grauer Literatur die zentrale sozialwissenschaftliche Literaturdatenbank für den deutschsprachigen Raum dar. Im Jahr 1980 eingerichtet und bis in das Jahr 1945 zurückreichend, verzeichnet sie zum Teil über Zulieferungen die im deutschsprachigen Raum erschienene sozialwissenschaftliche Forschungsliteratur.<sup>8</sup>

Die inhaltliche Auswertung der Literatur erfolgt in erster Linie in Form eines Kurzreferates sowie der Vergabe von inhaltlichen Schlagworten bzw. Deskriptoren sowie der Klassifikation.

Das Kurzreferat entspricht dem sog. informativen Referat (vgl. DIN-Norm 1426). Es gibt die zentralen Inhalte eines Dokumentes und damit sowohl die theoretischen und methodischen Ansätze als auch die wesentlichen Forschungsergebnisse wieder. Daneben wird neben dem behandelten Zeitraum auch der untersuchte geographische Raum aufgeführt (vgl. GESIS 2005: 2). Bei 40 Prozent der Dokumente bestehen diese Abstracts aus Autorenreferaten sowie Textteilen, die bei der Literatursichtung aus der Dokumentvor-

<sup>6</sup> Es gilt zu beachten, dass diese Form der Evaluation anhand intellektuell generierter Vergleichsdaten wie die inhaltliche Erschließung selbst von der eigenen Interpretation und damit von Vagheit und Unschärfe beeinflusst ist. Besondere Problembereiche, die bei dieser Form der Evaluation nur ungenügend berücksichtigt werden, bilden die mitunter sehr enge begriffliche Überschneidung bei den intellektuell und automatisch vergebenen Deskriptoren sowie die thematische Nähe mancher Klassifikationsnotationen.

<sup>7</sup> Für weitere Bewertungskriterien, etwa Dokument- bzw. Abstract-Art sowie Länge des Abstracts, die lediglich einen nachgeordneten Einfluss auf das Evaluationsergebnis hatten, siehe Kempf 2012.

<sup>8</sup> Den größten Anteil des Bestandes liefert GESIS – Leibniz-Institut für Sozialwissenschaften selbst. Neben Monographien und Sammelwerken, die über die Reihe A bestellt und in Autopsie erschlossen werden, werden zusätzlich knapp 300 Fachzeitschriften ausgewertet und auf Beitragsebene erschlossen. Daneben liefern Kooperationspartner, wie etwa das Institut für Arbeitsmarkt- und Berufsforschung (IAB) und das Wissenschaftszentrum Berlin für Sozialforschung (WZB) einen Teil des Bestandes.



lage übernommen werden. Gesamtaufnahmen von Sammelwerken zeichnen sich darüber hinaus durch die Aufnahme des Inhaltsverzeichnisses in das Indexat aus.

Die Schlagworte entstammen dem Thesaurus Sozialwissenschaften. Bestehend aus aktuell (Stand 2012) über 8.000 Deskriptoren sowie über 5.000 Synonymverweisen, den sog. Nicht-Deskriptoren, deckt er entsprechend des weit gefassten sozialwissenschaftlichen Fachbereichs eine große Bandbreite an Unterdisziplinen der Sozialwissenschaften ab.<sup>9</sup> Um den Umfang des Vokabulars überschaubar zu halten und sprachliche Veränderungen leichter darzustellen, wurde konsequent das Prinzip der Postkoordination verfolgt. Für Geographika, insbesondere außerhalb Europas, wurde ein sog. geographisches Upposting, bei dem automatisch entsprechende Oberbegriffe, wie etwa Kontinentbezeichnungen, hinzugefügt werden, eingeführt.

Die Klassifikation Sozialwissenschaften dient vorrangig der Zuordnung einer Publikation zu einem Wissenschaftsgebiet. Auf drei Hierarchieebenen verteilt, enthält sie 159 Klassifikationsnotationen aus etwa 20 Fachteilgebieten. Als Richtwert gilt die Vergabe von einer Haupt- und zwei Nebennotationen (vgl. ebd.).

## 6 Die Indexierungssoftware

Die beiden Testreihen wurden mit der Indexierungssoftware MindServer der Firma Recommind durchgeführt, die für die automatische Verschlagwortung und Klassifizierung von Dokumenteinheiten entwickelt wurde. Dabei stellt die Software ein sog. lernendes Verfahren dar. Anhand eines Trainingskorpus errechnet die Software Wahrscheinlichkeiten, nach denen einem neu zu erschließenden Dokument Kategorien in Form von Schlagworten und Klassifikationsnotationen zugeordnet werden. Die Software folgt somit in erster Linie einem statistischen Verfahrensansatz. Lediglich bestimmte Grundkomponenten, wie etwa Grundformreduktion und Derivation, basieren auf computerlinguistischen Analysen. Dadurch, dass der Software zur Repräsentation des Dokumentinhaltes entsprechend eines Additionsverfahrens ausschließlich Deskriptoren aus dem Thesaurus sowie die Klassifikation

Sozialwissenschaften zur Verfügung gestellt werden, wird das statistische Vorgehen zusätzlich um einen begriffsorientierten Verfahrensansatz erweitert.

Grundlage der Software bilden die sog. probabilistische latente semantische Indexierung sowie das Verfahren der Support Vector Machine. Erstere basiert auf einem patentierten Algorithmus, der sich wiederholende Themen und Konzepte identifiziert (vgl. Puzicha 2009, zitiert nach Keil/Tiesler 2010). Die Information über das Auftreten eines Terms in einem bestimmten Dokument wird gespeichert und für die weitere Indexierung genutzt. Diese Analyse wird um eine Variable erweitert, die die Wahrscheinlichkeit angibt, mit der ein Term oder ein Dokument zu einer bestimmten Klasse bzw. Thema oder Konzept gehört. Dieser Lernschritt wird für die Anordnung der Terme und Dokumente gleichsam in Form eines Vektorraums genutzt. Die sog. Support Vector Machine trennt als ein sog. Klassifikator diesen vieldimensionalen Vektorraum in Form von linearen Ebenen, um dadurch Klassenzugehörigkeiten zu bestimmen (vgl. Keil/Tiesler 2010).

Obgleich es sich auch bei dieser Form von semantischer Analyse lediglich um ein statistisches Verfahren handelt, geht dieses Verfahren über die Verarbeitung der einzelnen Wörter als Zeichenkette hinaus. So wird durch den Abgleich der einzelnen Dokumentdaten mit dem Gesamtbestand der Dokumentensammlung versucht, auch diejenigen Begriffe bzw. Begriffsinhalte zu erfassen, die lediglich latent in der Dokumentgrundlage enthalten sind.

## 7 Aufbau und Auswertung der Testreihen

Die Evaluation der Indexierungssoftware erfolgte in Form von zwei Testreihen von je zwei Testläufen auf der Basis einer festen Stichprobe von 280 Dokumenten. Entlang der beiden Testreihen wurden sowohl die Einstellungen der Software als auch die Korpora, auf deren Grundlage die Software trainiert wurde, verändert.

### 7.1 Die erste Testreihe

Für die erste Testreihe wurde der MindServer mit dem Gesamtbestand der Datenbank SOLIS bis Juli 2009 (ca. 368.000 Dokumente) trainiert.

Beim *ersten Testlauf* wurden die Standardeinstellungen des MindServer übernommen. Im Einzelnen bedeutete dies, dass zum einen keinerlei Einschränkung bei der Anzahl der von MindServer vergebenen Deskriptoren

<sup>9</sup> Die Sozialwissenschaften decken in Anlehnung an die UNESCO-Definition (1978) u. a. die Wissensgebiete und Anwendungsbereiche Soziologie, Politikwissenschaft, Erziehungswissenschaft und Kommunikationswissenschaft ab. Daneben wurden Bezeichnungen aus den Geistes- und Naturwissenschaften sowie Terme aufgenommen, die zum Gegenstandsbereich der Sozialwissenschaften zählen, auch wenn sie keine wissenschaftlichen Begriffe darstellen (vgl. IZ 2006).

und Notationen vorgenommen wurde. Zum anderen konnten sämtliche Kategorien (Deskriptoren/Notationen) automatisch vorgeschlagen werden, selbst wenn sie in den intellektuell generierten Trainingsdokumenten nur einmal vergeben wurden.<sup>10</sup> Daneben betrug das Verhältnis zwischen Recall und Precision +20,0, wodurch eine Indexierung zugunsten des Recall vorgenommen wurde.

Ziel des *zweiten Testlaufs* war es, die Präzision der automatischen Schlagwortvergabe zu erhöhen. Neben einer stärkeren Gewichtung der Precision um 20 Prozent bei der Deskriptorvergabe wurde daher sowohl für die Vergabe der Deskriptoren als auch der Notationen ein cut-off Level von zehn bzw. fünf eingesetzt. Ferner wurde, bezogen auf die Vergabe der Deskriptoren, die minimale Anzahl der Trainingsdokumente pro Kategorie auf 20 heraufgesetzt, wodurch sich die Gesamtzahl der berücksichtigten Deskriptoren leicht reduzierte.

## 7.2 Die zweite Testreihe

Mit dem Ziel, in Anlehnung an das Schalenmodell nach Krause (siehe oben) die Indexierungsleistung der Software in den Rand- mit jener in den Kernbereichen der Datenbank SOLIS zu vergleichen, wurden für die zweite Testreihe fachteilgebietsspezifische Versionen des MindServer aufgebaut. Hiermit war die Hypothese verbunden, durch die Eingrenzung der Trainingsmenge auf einzelne Fachteilgebiete die Ähnlichkeitserkennung zwischen den Dokumenten und damit die Erschließungsergebnisse weiter zu verbessern. Vorausgegangen war eine systematische fachteilgebietsspezifische Untersuchung der Indexierungsergebnisse der ersten Testreihe, die z. T. deutliche Unterschiede in der Indexierungsleistung zwischen den Kern- und Randbereichen der Datenbank erkennen ließ.<sup>11</sup>

Für die zweite Testreihe wurden daher anhand der intellektuell vergebenen Klassifikationsnotationen die Datenbestände zu den Kerngebieten Soziologie und Politikwissenschaft sowie dem exemplarischen Randgebiet der Geisteswissenschaften selektiert.<sup>12</sup> Für jedes dieser Fach-

teilgebiete wurde somit eine eigene Version des MindServer aufgebaut, die jeweils ausschließlich auf der Grundlage der entsprechenden selektierten Dokumentmenge trainiert wurde. Für die Soziologie umfasste diese ca. 142.000 Dokumente, für die Politikwissenschaft knapp 75.000 Dokumente und für die Geisteswissenschaften etwa 54.500 Dokumente. Analog wurden auch aus der Gesamtstichprobe diejenigen Dokumente selektiert, deren intellektuell vergebene Hauptnotation in die entsprechenden Klassifikationsbereiche fiel. Im Einzelnen waren dies 56 Dokumente für das Fachteilgebiet Soziologie, 68 Dokumente für den Bereich Politikwissenschaft und 40 Dokumente für den Bereich Geisteswissenschaften.

Für den *dritten Testlauf* wurden wie bereits für den ersten Testlauf die Standardeinstellungen des MindServer verwendet. Für den *vierten Testlauf* wurden zunächst die gleichen Einstellungen wie für den zweiten Testlauf vorgenommen. Verbunden mit den deutlich schmaleren Trainingskorpora aufgrund der vorausgegangenen fachteilgebietsspezifischen Selektion, führte dies allerdings zu keinen befriedigenden Ergebnissen.<sup>13</sup> Für die Schlagwortvergabe wurde daher ein ausgeglichenes Verhältnis zwischen Recall und Precision ausgewählt. Für die Vergabe der Notation entsprach dieses, wie in den vorausgegangenen Testläufen, den Standardeinstellungen.

## 7.3 Die Testergebnisse im Vergleich

In der ersten Testreihe ergeben die Veränderung des Verhältnisses zwischen Recall und Precision und die Einführung eines cut-off-Niveaus von zehn im zweiten Testlauf für die Schlagwortvergabe eine signifikante Erhöhung des Precision-Wertes von 32 Prozent auf 53 Prozent bei einem Rückgang des Recall von gut 40 Prozent auf 30 Prozent.<sup>14</sup> Die Indexierungskonsistenz nimmt ebenfalls

<sup>10</sup> Die maximale Anzahl an Dokumenten in der Trainingsmenge pro Kategorie betrug hingegen bei allen vier Testläufen 25.000.

<sup>11</sup> In den Tab. 2 bis 4 sind zu einem besseren Gesamtvergleich der Indexierungsleistung daher auch für die erste Testreihe die Ergebnisse differenziert nach sowohl Kernbereich Soziologie und Politikwissenschaft als auch dem exemplarischen Randbereich Geisteswissenschaften aufgeführt.

<sup>12</sup> Die Auswahl des Randbereichs Geisteswissenschaften hing mit der Zusammensetzung der Gesamtstichprobe zusammen. So stellten die Geisteswissenschaften in der Stichprobe einen der größeren Datenbestände zu den Randgebieten dar.

<sup>13</sup> Aufgrund der stärkeren Gewichtung der Precision um 20 Prozent lieferte die Indexierungssoftware im Durchschnitt lediglich zwischen einem und zwei Deskriptoren. Bei diesen Deskriptoren handelte es sich vor allem um Bezeichnungen, die in den Trainingsmengen aufgrund des geographischen Uppostings sehr häufig vergeben wurden (z. B. „Bundesrepublik Deutschland“). Der Ähnlichkeitsabgleich mit anderen Dokumenten erfolgte in einem zu eng gefassten Bezugsraum.

<sup>14</sup> Vergleichswerte aus der Fachinformation für die Psychologie liegen bei einer durchschnittlichen Übereinstimmung der automatisch generierten Deskriptorvorschläge mit der manuellen Indexierung von 46,7 Prozent (Recall). Bezogen auf die durchschnittliche Gesamtzahl der automatisch vergebenen Deskriptoren bedeutete dies eine Übereinstimmung mit dem intellektuell generierten Indexat um 35,4 Prozent (Precision) (vgl. Gerards et al. 2006 sowie Gerards 2011). In der wirtschaftswissenschaftlichen Fachinformation liegt



leicht zu. Auch bei der Vergabe der Klassifikationsnotation ist ein leichter Anstieg des Precision-Wertes sowie der Indexierungskonsistenz zu verzeichnen (siehe Tab. 1).

**Tabelle 1:** Evaluationsergebnis der ersten beiden Testläufe für Deskriptor- und Notationsvergabe.

	Deskriptorvergabe		Notationsvergabe	
	TL I (n = 271) <sup>15</sup>	TL II (n = 263)	TL I (n = 262)	TL II (n = 254)
Recall	44,0 % (48,6 %) <sup>16</sup>	30,9 % (36,4 %)	63,0 % (70,8 %)	58,3 % (79,2 %)
Precision	32,4 % (35,8 %)	53,0 % (62,3 %)	37,5 % (42,5 %)	40,0 % (54,3 %)
Indexierungs- konsistenz	37,3 % (41,3 %)	39,0 % (46,0 %)	45,2 % (52,0 %)	46,4 % (61,8 %)

Der Aufbau fachteilgebietsspezifischer MindServer-Versionen für die zweite Testreihe führt zu einer deutlichen Erhöhung von Precision und Indexierungskonsistenz sowohl für die Deskriptor- als auch die Notationsvergabe (siehe Tab. 2 u. 3). Dies wird aus dem Vergleich der beiden Testläufe I und III, für die beide die Standard-einstellungen der Indexierungssoftware verwendet wurden, deutlich. Verbunden mit der relativen Erhöhung des Precision-Wertes bei den Systemeinstellungen für die Vergabe der Deskriptoren beim vierten Testlauf, lässt sich die Präzision der automatisch generierten Indexierungsergebnisse weiter erhöhen. Der durchweg höhere Precision-Wert für die Politikwissenschaft könnte mit einem klarer strukturierten Begriffsapparat sowie weniger häufig verwendeten Allgemeinbegriffen in diesem Bereich des Thesaurus Sozialwissenschaften verbunden sein. Auch bei der Indexierungskonsistenz zeichnet sich für alle drei Fachteilgebiete durch die fachspezifischen Trainingsmengen eine Erhöhung ab.

die Indexierungskonsistenz bei der Schlagwortvergabe bei 36 Prozent (Groß 2010: 1130). Obgleich die Einrichtungen z. T. dieselbe Indexierungssoftware verwenden, sind die Anwendungsszenarien allerdings derart unterschiedlich, dass ein direkter Vergleich der Indexierungsergebnisse nicht möglich ist.

<sup>15</sup> Aufgrund der vorgenommenen Systemeinstellungen für die verschiedenen Testläufe, dazu gehörte auch die minimale Textlänge aus Titel und Abstract bzw. Autorenreferat, die für die Gesamtbetrachtung der Ergebnisse allerdings nur gering ins Gewicht fiel, variierte die Gesamtzahl der Stichprobendokumente mehrfach.

<sup>16</sup> Die Recall- und Precision-Werte in Klammern würden erzielt, wenn die durchschnittliche Anzahl an Deskriptoren einberechnet würde, die zwar nicht im intellektuell generierten Indexat vergeben bei der Evaluation jedoch als durchaus zutreffend und inhaltstragend bewertet wurden (vgl. Fußnote 6).

**Tabelle 2:** Evaluationsergebnis aller vier Testläufe für die Deskriptorvergabe.

	TL I	TL II <sup>17</sup>	TL III	TL IV
<b>Soziologie</b> (n = 56)				
Recall	46,6 %	32,5 %	45,1 %	42,1 %
Precision	31,5 %	51,4 %	37,4 %	41,1 %
Ind.-Kons.	37,6 %	39,9 %	40,9 %	41,6 %
<b>Politik</b> (n = 68)				
Recall	46,0 %	32,2 %	43,5 %	41,5 %
Precision	36,1 %	60,1 %	42,4 %	46,4 %
Ind.-Kons.	40,0 %	41,9 %	42,9 %	43,2 %
<b>Geistesw.</b> (n = 40)				
Recall	42,0 %	30,0 %	38,3 %	39,5 %
Precision	32,0 %	51,4 %	36,0 %	41,0 %
Ind.-Kons.	36,3 %	37,9 %	37,1 %	41,7 %

In der Gesamtschau aller vier Testläufe stechen für die Schlagwortvergabe sowohl die Recall- als auch die Precision-Werte des zweiten Testlaufs deutlich hervor. Verbunden mit der stärkeren Gewichtung der Precision im Vergleich zum Recall um 20 Prozent und der Einführung eines cut-off Levels von zehn, erscheinen hier die Indexierungsergebnisse am nächsten an der intellektuellen Erschließung. Die starke Erhöhung der Präzision geht allerdings sehr deutlich auf Kosten des Recall.

Ein genauer Blick auf die Notationsvergabe zeigt ferner, dass sich mit dem Aufbau fachteilgebietsspezifischer MindServer-Versionen für alle drei Fachteilgebiete eine verbesserte Platzierung der intellektuell vergebenen Hauptnotation erzielen lässt (siehe Tab. 4). Am deutlichsten

**Tabelle 3:** Evaluationsergebnis aller vier Testläufe für die Notationsvergabe.

	TL I	TL II	TL III	TL IV
<b>Soziologie</b> (n = 56)				
Recall	59,1 %	59,1 %	56,9 %	41,8 %
Precision	30,9 %	36,8 %	33,3 %	49,5 %
Ind.-Kons.	38,9 %	43,2 %	40,0 %	42,2 %
<b>Politik</b> (n = 68)				
Recall	79,8 %	74,2 %	71,7 %	59,7 %
Precision	46,9 %	48,1 %	58,4 %	73,9 %
Ind.-Kons.	59,0 %	58,3 %	64,4 %	66,0 %
<b>Geistesw.</b> (n = 40)				
Recall	45,2 %	45,2 %	56,1 %	36,0 %
Precision	27,2 %	31,4 %	37,3 %	43,9 %
Ind.-Kons.	33,9 %	37,1 %	44,8 %	40,5 %

<sup>17</sup> Die Kursiv-Setzung soll an dieser Stelle die spezifischen Systemeinstellungen des zweiten Testlaufs in Form des verwendeten cut-off-Llevels zum Ausdruck bringen. Er ist daher nur eingeschränkt mit den anderen Testläufen vergleichbar. Die Unterstreichung (Testläufe III u. IV) dient dazu auszudrücken, dass diesen Testläufen fachteilgebietsspezifische Trainingskorpora zugrunde liegen.

**Tabelle 4:** Evaluationsergebnis aller vier Testläufe für Vergabe und Ranking-Position der Hauptnotation.

	TL I	TL II	TL III	TL IV
<b>Soziologie (n = 56)</b>				
Hauptklass. gef.	78,2 %	78,4 %	85,2 %	69,2 %
Ranking-Position	1,7	1,6	1,4	1,2
<b>Politik (n = 68)</b>				
Hauptklass. gef.	92,4 %	87,3 %	89,4 %	87,8 %
Ranking-Position	1,7	1,5	1,2	1,4
<b>Geistesw. (n = 40)</b>				
Hauptklass. gef.	63,2 %	59,0 %	77,5 %	57,9 %
Ranking-Position	2,7	1,7	1,4	1,1

ten zeigt sich diese im Randbereich der Datenbank. Hier lässt sich bereits mit der Einführung eines cut-off Levels, wie sie beim zweiten Testlauf auch für die Notationsvergabe erfolgte, eine deutlich höhere Ranking-Position der intellektuell vergebenen Hauptnotation erzielen. Daneben lässt sich mit dem Aufbau fachteilgebietsspezifischer Versionen für den Randbereich auch die Häufigkeit, mit der die intellektuell vergebene Hauptnotation auch automatisch zugeordnet wird, deutlich erhöhen. Für die Kernbereiche trifft dies nur eingeschränkt zu. Gleichwohl bleibt das hohe Niveau, auf dem die intellektuell vergebene Hauptnotation von der Indexierungssoftware generiert wird, bestehen. Hierbei ist zu beachten, dass sich die Kernbereiche, insbesondere im Fall der Soziologie, auf deutlich mehr Notationen der Klassifikation Sozialwissenschaften verteilen, als dies für den Randbereich der Geisteswissenschaften zutrifft.

## 8 Fazit und Ausblick

Als zentrales Ergebnis der vorliegenden Evaluation kann festgehalten werden, dass – Deskriptor- und Notationsvergabe gemeinsam betrachtet – die automatisch generierten Indexierungsergebnisse für die Kerngebiete der Datenbank SOLIS tendenziell eine höhere Übereinstimmung mit dem intellektuellen Indexat aufweisen als die Ergebnisse für die Randgebiete. Eine automatische Vorindexierung für die Randgebiete, wie sie von Krause vorgeschlagen wird, würde unter diesen Testbedingungen im Vergleich zu den Kerngebieten weniger vollständige Indexierungsvorschläge liefern. Dies erklärt sich aus der niedrigeren Gesamtzahl an Trainingsdokumenten zu den Randbereichen in SOLIS, wodurch die Kontext- bzw. Ähnlichkeitserkennung zu anderen Dokumenten erschwert wird. Die entlang der Testläufe insgesamt relativ heterogene Ergebnislage zwischen Kern- und Randgebiet lässt allerdings eine genauere fachteilgebietsspezifische

Untersuchung für andere Randbereiche der Datenbank notwendig erscheinen. Hierbei ließe sich darauf zurückgreifen, dass bereits aktuell eine Vorindexierung durch die Software MindServer technisch möglich ist. Diese Indexierungsergebnisse ließen sich differenziert nach den einzelnen intellektuell vergebenen Hauptnotationen auswerten.

Ferner geht aus der Evaluationsstudie hervor, dass sich durch die Einführung eines cut-off Levels die Präzision der automatisch generierten Indexierungsvorschläge deutlich erhöhen lässt. Dies gilt vor allem für die Vergabe der Deskriptoren. Beim Einsatz einer semiautomatischen Indexierung für die Randbereiche von SOLIS sollte somit – auch um für die Indexierer den zeitlichen Aufwand bei der Auswertung des automatisch generierten Indexates gering zu halten – die Anzahl der automatisch vergebenen Deskriptoren eingeschränkt werden.<sup>18</sup>

Der Aufbau fachteilgebietsspezifischer MindServer-Versionen führt ebenso sowohl für die Deskriptor- als auch für die Notationsvergabe zu einer Präzisionserhöhung der Indexierungsergebnisse. Die Erhöhung der Präzision, die durch die größere Homogenität der Trainingsmengen erzielt wird und die ihrerseits die Ähnlichkeitserkennung zu anderen Dokumenten erleichtert, bleibt jedoch hinter dem Precision-Anstieg zurück, wie er über die Einführung eines cut-off Levels im zweiten Testlauf erzielt wird.

Mit Blick auf die praktische Umsetzung einer automatischen Indexierung zeichnet sich aus der Zusammenschau der durchgeführten Testläufe ab, dass der Aufwand, fachteilgebietsspezifische Versionen der Indexierungssoftware aufzubauen, nicht im Verhältnis zum zu erwartenden Qualitätsgewinn stehen würde. Nicht nur ist die dadurch erzielte Erhöhung der Indexierungspräzision geringer als bei der Festlegung eines cut-off Levels und einer gleichzeitigen Erhöhung des Precision-Wertes in den Systemeinstellungen, auch die Vorselektion nach Fachteilgebieten und die Pflege derartiger Versionen der Indexierungssoftware ließen sich nur schwer ohne größeren Aufwand in den alltäglichen Geschäftsgang integrieren.

Als Ausblick ließe sich das Verhalten der Indexierungssoftware in Bezug auf bestimmte Erschließungsrichtlinien näher untersuchen, um zu einer noch differenzierteren Bewertung der automatisch generierten

<sup>18</sup> Für die Einschränkung der automatisch generierten Klassifikationsnotationen, so ging aus zusätzlichen explorativen Auswertungen hervor, ließe sich mit Abstandsmessungen der Konfidenzwerte arbeiten. Liegen die Konfidenzwerte der automatisch vorgeschlagenen Klassifikationsnotationen verhältnismäßig weit auseinander, lässt sich davon ausgehen, dass die nachfolgend vorgeschlagenen Klassifikationsnotationen in den meisten Fällen nicht intellektuell vergeben würden.



Indexierungsergebnisse zu gelangen. So wäre zu klären, inwiefern vor dem Hintergrund der spezifischen Erschließungsrichtlinien mit einem bestimmten cut-off Level die inhaltliche Wiedergabe der dokumentarischen Bezugseinheiten gelingt. Denkbar ist etwa, dass in Verbindung mit dem geographischen Uposting vor allem eine geographische Einordnung vorgenommen wird. Schließlich könnte das Suchverhalten der Nutzer noch stärker einbezogen werden. So ließe sich durch die Installation einer Tagging-Funktion Aufschluss darüber erhalten, welche Bezeichnungen Nutzer wählen, um Dokumentinhalte wiederzugeben und aufzufinden.

## Literatur

- Bertram, Jutta (2005): Einführung in die inhaltliche Erschließung. Grundlagen – Methoden – Instrumente. Würzburg: ERGON-Verlag.
- Deutsche Nationalbibliothek (DNB) (2010): PETRUS. Prozessunterstützende Software für die digitale Deutsche Nationalbibliothek. Interne Präsentation, Mai 2010.
- Gerards, Michael/Gerards, Andreas/Weiland, Peter (2006): Der Einsatz der automatischen Indexierungssoftware AUTINDEX im Zentrum für Psychologische Information und Dokumentation (ZPID) <http://www.zpid.de/download/PSYNDEXmaterial/autindex.pdf> [02.04.2011].
- Gerards, Michael (2011): Semiautomatische Erschließung von Psychologie-Information. Präsentation im Rahmen des PETRUS-Workshops an der Deutschen Nationalbibliothek Frankfurt/Main, 21./22. März 2011 [http://files.d-nb.de/pdf/petrus/semiautomatische\\_erschliessung\\_zpid.pdf](http://files.d-nb.de/pdf/petrus/semiautomatische_erschliessung_zpid.pdf) [22.05.2011].
- GESIS – Informationszentrum Sozialwissenschaften (Hg.) (2005): Regelwerk für die Literaturdokumentation Sozialwissenschaften. Regeln für die inhaltliche Erschließung sozialwissenschaftlicher Literatur.
- Groß, Thomas (2010): Die Implementierung eines automatischen Indexierungsverfahrens am Beispiel der Deutschen Zentralbibliothek für Wirtschaftswissenschaften. Masterarbeit im Rahmen des postgradualen Fernstudiums Master of Arts. <http://edoc.hu-berlin.de/master/gross-thomas-2010-05-08/PDF/gross.pdf> [10.11.2010].
- Groß, Thomas/Faden, Manfred (2010): Automatische Indexierung elektronischer Dokumente an der Deutschen Zentralbibliothek für Wirtschaftswissenschaften. In: Bibliotheksdienst, 44. Jg., 12, 1120–1135.
- Informationszentrum Sozialwissenschaften (IZ) (2006): Thesaurus Sozialwissenschaften. Alphabetischer Teil/Systematischer Teil. Bonn: Informationszentrum Sozialwissenschaften.
- KASCADE – [http://www.uni-duesseldorf.de/projekte/kascade/kas\\_home](http://www.uni-duesseldorf.de/projekte/kascade/kas_home) [06.04.2011].
- Keil, Stefan/Tiesler, Philipp et al. (2010): Automatische Erschließung für die Datenbank SOLIS. Internes Arbeitspapier von GESIS – Leibniz-Institut für Sozialwissenschaften.
- Kempf, Andreas Oskar (2012): Automatische Indexierung in der sozialwissenschaftlichen Fachinformation. Eine Evaluationsstudie zur maschinellen Erschließung für die Datenbank SOLIS. Berliner Handreichungen zur Bibliotheks- und Informationswissenschaft, 329. Berlin: Institut für Bibliotheks- und Informationswissenschaft der Humboldt-Universität zu Berlin.
- Krause, Jürgen (1996): Informationserschließung und -bereitstellung zwischen Deregulation, Kommerzialisierung und weltweiter Vernetzung – Schalenmodell. IZ-Arbeitsbericht Nr. 6, Bonn: Informationszentrum Sozialwissenschaften.
- Krause, Jürgen (2006): Shell Model, Semantic Web and Web Information Retrieval. In: Harms, Ilse/Luckhardt, Heinz-Dirk/Giesen, Hans W. (Hg.) Information und Sprache. Beiträge zu Informationswissenschaft, Computerlinguistik, Bibliothekswesen und verwandten Fächern. Festschrift für Harald H. Zimmermann. K. G. Saur: München, 95–106.
- Lingelbach-Hupfauer, Carmen/Laute, Hartwig (2009): Die semi-automatische Indexierung von Zeitungsartikeln. In: Info 7, Jg. 24, Heft 2, 48–50.
- Lingelbach-Hupfauer, Carmen (2011): Die semi-automatische Erschließung von Zeitungs- und Zeitschriftenartikeln in der Presseudokumentation des ZDF. Präsentation im Rahmen des PETRUS-Workshops an der Deutschen Nationalbibliothek Frankfurt/Main, 21./22. März 2011 [http://files.d-nb.de/pdf/petrus/semi-automatische\\_indexierung\\_zdf.pdf](http://files.d-nb.de/pdf/petrus/semi-automatische_indexierung_zdf.pdf) [19.05.2011].
- Mayr, Philipp/Mutschke, Peter/Schaer, Philipp/Sure, York (2009): Mehrwertdienste für das Information Retrieval: das Projekt IRM. <http://www.ib.hu-berlin.de/~mayr/arbeiten/IRM-ISK009.pdf> [03.04.2011].
- MILOS – [http://www.ub.uni-duesseldorf.de/projekte/milos/mil\\_home](http://www.ub.uni-duesseldorf.de/projekte/milos/mil_home) [05.04.2011].
- Nohr, Holger (2004<sup>5</sup>): Theorie des Information Retrieval II: Automatische Indexierung, in: Kuhlen, Rainer et al. (Hg.) Grundlagen der praktischen Information und Dokumentation, München: Saur, 215–225.
- Nohr, Holger (2005<sup>3</sup>): Grundlagen der automatischen Indexierung – Ein Lehrbuch. Berlin: Logos-Verlag.
- Normausschuss Bibliotheks- und Dokumentationswesen (1988): DIN 1426. Inhaltsangaben von Dokumenten; Kurzreferate, Literaturberichte. Berlin u. a.: Beuth.
- Organisation der Vereinten Nationen für Erziehung, Wissenschaft und Kultur – Büro für internationale Normen und Rechtsfragen (Hg.) (1979): Empfehlungen zur internationalen Vereinheitlichung der Statistiken über Wissenschaft und Technologie der UNESCO-Generalkonferenz vom 28.11.1978.
- Puzicha, Jan (2009): Informationen finden! Intelligente Suchmaschinenteknologie & automatische Kategorisierung. Technical Whitepaper – Grundlagen der Informationsgewinnung. Mind-Server. Publikation der Firma Recommind.
- Rolling, L. (1981): Indexing consistency, quality and efficiency. In: Information Processing & Management, 17. Jg., 69–76.
- Schöning-Walter, Christa (2011): Automatische Erschließungsverfahren für Netzpublikationen. Stand der Arbeiten im Projekt PETRUS. Präsentation bei einem Kolloquium von GESIS – Leibniz-Institut für Sozialwissenschaften, Februar 2011.
- Siegmüller, Renate (2007): Verfahren der automatischen Indexierung in bibliotheksbezogenen Anwendungen. In: Berliner Handreichungen zur Bibliotheks- und Informationswissenschaft, Heft 214. Berlin. <http://www.ib.hu-berlin.de/~kumlau/handreichungen/h214/h214.pdf> [03.04.2011].
- Stock, Wolfgang G./Stock, Mechtild (2008): Wissensrepräsentation. Informationen auswerten und bereitstellen. München: Oldenbourg Verlag.

- Stubbs, Edgardo/Mangiaterra, Norma E./Martinez, Anna M. (1999): Internal quality audit of indexing: A new application of interindexer consistency. *Cataloguing & Classification Quarterly*, Jg. 28, Heft 4: 53–69.
- Support Vector Machine – Wikipedia [http://de.wikipedia.org/wiki/Support\\_Vector\\_Machine](http://de.wikipedia.org/wiki/Support_Vector_Machine) [09.04.2011].
- Wissel, Verena (2011): Erfahrungsbericht und Schlussfolgerungen des DIPF zur Erprobung automatischer Erschließung. Präsentation im Rahmen des PETRUS-Workshops an der Deutschen Nationalbibliothek Frankfurt am Main, 21./22. März 2011. [http://files.d-nb.de/pdf/petrus/dipf\\_erfahrungsbericht.pdf](http://files.d-nb.de/pdf/petrus/dipf_erfahrungsbericht.pdf) [15.05.2011].
- Womser-Hacker, Christa (2004<sup>5</sup>): Theorie des Information Retrieval III: Evaluierung, in: Kuhlen, Rainer/Seeger, Thomas/Strauch, Dietmar (Hg.) *Grundlagen der praktischen Information und Dokumentation*. München: K. G. Saur, 227–235.

## Danksagungen

Mein besonderer Dank gilt meinen KollegInnen Monika Zimmer, Hannelore Schott und Jan Hendrik Schulz, die mich beim Aufbau und bei der Durchführung sowie zum

Teil bei der Auswertung der Testläufe unterstützt haben. Bei Philipp Schaer und Nadine Dulisch möchte ich mich vor allem für die technische Unterstützung bedanken.



**Dr. Andreas Oskar Kempf**

GESIS – Leibniz-Institut für Sozialwissenschaften  
Fachinformation Sozialwissenschaften  
Unter Sachsenhausen 6–8  
50667 Köln  
[andreas.kempf@gesis.org](mailto:andreas.kempf@gesis.org)  
[www.gesis.org](http://www.gesis.org)

Dr. Andreas Oskar Kempf ist Absolvent des weiterbildenden Masterstudiengangs Bibliotheks- und Informationswissenschaft an der Humboldt-Universität zu Berlin. Nach dem Studium der Kulturwissenschaften und einer Promotion im Fach Soziologie ist er aktuell wissenschaftlicher Mitarbeiter bei GESIS – Leibniz-Institut für Sozialwissenschaften. Verantwortlich für den Thesaurus und die Klassifikation Sozialwissenschaften zählen zu seinen Forschungsgebieten neben der automatischen Indexierung die Stärkung der Interoperabilität sowie die Einbindung der Erschließungsinstrumente in Anwendungen des Semantic Web.